IES Grant: R305C120001

JOURNAL OF RESEARCH ON TECHNOLOGY IN EDUCATION
https://doi.org/10.1080/15391523.2020.1826022

The impact of conversational agents' language on summary writing

Haiying Li[a] and Art C. Graesser[b]

[a]The Office of Enrollment Research & Analytics, Iowa State University, Ames, Iowa, USA; [b]Department of Psychology and Institute for Intelligent Systems, University of Memphis, Memphis, Tennessee, USA

ABSTRACT

This study investigated how computer agents' language style affects summary writing in an Intelligent Tutoring System, called CSAL AutoTutor. Participants interacted with two computer agents in one of three language styles: (1) a *formal* language style, (2) an *informal* language style, and (3) a *mixed* language style. Primary results indicated that participants improved the quality of summary writing, spent less time writing summaries, and had lower syntactic complexity but more non-narrative summaries on posttest than pretest. However, this difference was not affected by the discourse formality that agents used during instruction. Results also showed participants rated peer summaries more accurately for cause/effect texts in the formal and mixed conditions, but generated summaries with lower referential cohesion in the informal condition on posttest than pretest.

ARTICLE HISTORY
Received: 1 November 2019
Revised 7 September 2020
Accepted 16 September 2020

KEYWORDS
Formality; summary writing; academic language; conversational language; agent

Introduction

Language is one of the most powerful tools that teachers use to organize and implement instructional activities to help students achieve learning goals (Denton, 2013). The teachers' choice of language, discourse, and higher-level communication patterns (hereafter called language) is related to both the success of student learning (e.g., Hebert et al., 2016; Kalinowski et al., 2019) and the development of students' language (Fillmore & Snow, 2003; Gámez & Lesaux, 2012, 2015; Lucero, 2014). Prior studies on teacher language contrasted academic language and conversational language. We defined academic language as the pre-planned, well-organized, and coherent language used for academic communication, either spoken or written, with comparatively low reliance on the contexts or shared common grounds of conversational participants (Clark, 1996). Conversational language, opposite of academic language, was defined as the spontaneous, less-organized, and more disjointed discourse, either spoken or written, with much reliance on the contexts and common ground shared by the speaker/writer and the audience.

Most studies on teacher language concentrate on its theoretical frameworks (e.g., Snow & Uccelli, 2009). Other studies are confined to correlational research that investigates the relationship between teacher language and learning (Gámez & Lesaux, 2012, 2015). No causal studies, to date, have been conducted on teacher language due to the difficulty in consistently manipulating teacher language in traditional classroom settings.

To address this challenge, some researchers have designed conversational agents to manipulate the language of computer agents and investigate the impact of computer agent language on learning in Intelligent Tutoring Systems (ITSs) (Li & Graesser, 2017, 2020), multimedia lessons in Chinese (Lin et al., 2020) and German (Reichelt et al., 2014) languages, a massive online open

CONTACT Haiying Li haiyingl@iastate.edu 361 Carver, 411 Morrill Rd, Ames, IA 50011, USA.

IES Grant: R305C120001

course (MOOC) environment (Riehemann & Jucks, 2018), or educational games (Moreno & Mayer, 2000, 2004). However, these studies manipulated agent language differently than the focus of this study on language formality. Most previous studies manipulated agent language using personal pronouns (Lin et al., 2020; Moreno & Mayer, 2000, 2004; Reichelt et al., 2014; Riehemann & Jucks, 2018). Only a few studies considered multiple-textual levels to manipulate agent language, including word, syntax, cohesion, and genre (Li & Graesser, 2017, 2020). Prior studies on teacher/agent language primarily investigated the effect of agent language on learning. No studies, to date, have examined the impact of teacher/agent language on both learning and language use, which would be very helpful for teachers/agents in delivering more efficient lectures and instructions.

This study fills gaps in prior research with the goal to investigate how agent language affects learning and language use. We investigated how agent language that considers multiple levels of words, syntax, cohesion, and genres influences learning of summary writing and language use in written summaries. Learning was measured by quality of summaries, accuracy of self-ratings, and accuracy of peer ratings. Language use was measured by language formality and its five underlying components, including word abstractness, syntactic complexity, referential cohesion, deep cohesion, and non-narrativity. These multiple types of evaluation allow for more comprehensively unpacking the effects of agent language on participants' learning and language use.

Teacher language

The early line of research on teacher language concentrated on theories and theoretical frameworks. The pragmatics-based framework (Snow & Uccelli, 2009) provides a comprehensive view of teacher language and proposes measures for each level of language, including linguistic features (e.g., interpersonal stance, information load, organization of information, lexical choices, representational congruence) and cognitive features (genre mastery, command of reasoning/argumentative strategies, disciplinary knowledge).

Early researchers conducted empirical studies on particular levels of teacher language using this pragmatics-based framework, such as at the lexical and/or syntactic levels (Galloway & Uccelli, 2015; Gámez & Lesaux, 2012, 2015). For example, Gámez and Lesaux (2012, 2015) reported a significant correlation of teachers' use of sophisticated, academic vocabulary and complexity of syntax with students' reading comprehension performance or vocabulary skills when controlling for classroom, school, and students' performance at the start of the year. These researchers measured syntactic complexity with embedded clauses, which was manually coded (Gámez & Lesaux, 2012) and very costly and time-consuming to collect. The lack of an automated tool to extract complex linguistic features is likely a primary reason for early research in this area being constrained to lexical and syntactic levels.

Agent language

Advances in educational technologies allow us to investigate causal relationships between teacher language and student learning and/or language use. Specifically, the ITS and educational games allow for the design of conversational computer agents to simulate human teachers with different language features and thereby conduct causal studies by manipulating agent language (Li & Graesser, 2017, 2020; Lin et al., 2020; Moreno & Mayer, 2000, 2004; Reichelt et al., 2014; Riehemann & Jucks, 2018). The fields of computational linguistics and natural language processing (NLP) provide automated language analysis tools that extract the language features at the multiple levels that were identified in the framework on teacher language (Graesser, McNamara, et al., 2014).

Moreno and Mayer (2000, 2004) designed an agent-based multimedia educational game, in which they manipulated an on-screen agent language into personalized speech (e.g., *I* and *you*) versus non-personalized speech (e.g., 3rd-person). They found that college students who received personalized agent messages performed better on science learning, measured by retention tests and problem-solving transfer tests. Moreover, they found that students in the personalized speech condition reported less difficulty in using the program. Findings from these studies have laid the foundation for the personalization theory in the design of conversation-based, computer-assisted learning environments. These findings about learning outcomes were further supported by a meta-analysis with 74 empirical studies from 1981 to 2012 (Ginns et al., 2013): textual materials in a conversational style showed positive small-to-medium effects on retention and a medium effect of transfer, but not on perceived interest.

Some studies, however, reported more complex results. For example, Reichelt et al. (2014) found a positive effect of conversational style on retention, but not on transfer on the subject matter of psychology. The same results were found in a study on the instruction of the human cardiovascular system (Lin et al., 2020). However, Riehemann and Jucks (2018) used the same lesson materials from psychology and found a conversational language style benefited transfer. The reason for inconsistent findings is likely due to different settings where experiments were conducted, the former in a research lab and the latter in a MOOC environment. These studies are restricted by the agent language measure, which used personal pronouns to represent the personalized speech of the agent.

Li & Graesser (2017, 2020) recently designed a conversation-based ITS to examine the agent language at multiple levels of language. They introduced Coh-Metrix *formality* scores, a composite measure of text difficulty (Graesser, McNamara, et al., 2014), to investigate agent language. The formality score has five major components of discourse: words, syntax, referential cohesion, deep cohesion, and genre. Academic language and conversational language are at two extreme ends of the formality continuum, where academic language is at one end (i.e., formal language) and conversational language at the other (i.e., informal language).

The Coh-Metrix formality scores are computed with the average scores of five primary components that are extracted by the Coh-Metrix tool (cohmetrix.com; Graesser, McNamara, et al., 2014). The Coh-Metrix text analysis tool was developed based on the multilevel theoretical framework, including textbase, situation model, genre, rhetorical structure, and pragmatic communication (Graesser & McNamara, 2011), most of which are similar to those listed in the pragmatics-based framework (Snow & Uccelli, 2009). The Coh-Metrix five components were extracted using a principal components analysis based on a corpus of texts that individuals are exposed to from kindergarten to early years of college:

Word abstractness is the inverse of the *word concreteness* component generated by Coh-Metrix. Texts are easier to process if they contain content words that are concrete, meaningful, and evoke mental images compared to abstract words that lack visual representations in the mind.

Syntactic complexity is the inverse of the *syntactic simplicity* component generated by Coh-Metrix. The high occurrence of left-embedded syntax and noun-phrase density (i.e., many words in a noun-phrase) increases the complexity of sentences, which is challenging and difficult to process.

Referential cohesion refers to words and ideas that overlap across explicit sentences and in the entire text. This overlap forms explicit threads that connect the text for the reader.

Deep cohesion refers to causal, intentional, or other types of connectives or conceptual ideas that help the reader form a more coherent, explicit, and deeper understanding of the text at the level of the causal situation model.

Non-Narrativity is the inverse of the *narrativity* component generated by Coh-Metrix. Narrative texts tell a story, with characters, events, places, and things that are familiar to readers or listeners. It is closely affiliated with everyday oral conversation.

IES Grant: R305C120001

Academic or formal language increases with high word abstractness, syntactic complexity, referential cohesion, deep cohesion, and non-narrativity. Conversely, conversational language or informal language decreases with low word abstractness, syntactic complexity, referential cohesion, deep cohesion, and non-narrativity. The Coh-Metrix formality scores, computed by the average of these five standardized component scores, with 0 as the medium formality ($M = 0$), equal to the formality of science textbooks at grade 6 (Graesser, McNamara, et al., 2014). The higher numbers above 0 represent more formal language, whereas the lower scores below 0 represent more informal language (Graesser, McNamara, et al., 2014).

Li and Graesser (2017, 2020) used this multi-level formality analysis to manipulate agent language. They reported that the agents' informal discourse yielded higher performance on the quality of summary writing, consistent with prior work showing that agents' informal discourse enhanced learning (Ginns et al., 2013; Lin et al., 2020; Moreno & Mayer, 2000, 2004; Reichelt et al., 2014; Riehemann & Jucks, 2018). They also found that agents' informal discourse elicited higher reports of text difficulty, which were inconsistent with prior findings that the agents' informal language elicited less difficulty reported on programs (Moreno & Mayer, 2004). This inconsistency can potentially be attributed to differences in subject matters (comprehension vs. science), measures of learning outcomes (summary writing vs. constructed responses), learning environments (ITS vs. educational game), report content (reading text vs. program of educational game), and different measures of agent language (formality at multi-textual levels vs. personal pronouns).

Even though Li and Graesser (2017, 2020) advanced the research on agent language by using the multilevel measures of formality, those studies were limited to the measure of learning outcomes using the quality of summary writing alone. This present study developed this line of research further by exploring two research questions:

1. Does conversational agents' formality during instruction affect learning of text structures?
2. Does the conversational agents' formality during instruction affect participants' language use when they write summaries?

To answer the first question, the study included more measures of learning outcomes than prior studies, which were the quality of summaries, self-rating of summaries, peer-rating of summaries, and writing time. To answer the second question, the study measured language use in participants' written summaries at multiple levels, including formality as well as its underlying five primary components: word abstractness, syntactic complexity, referential cohesion, deep cohesion, and non-narrativity.

Methods

Participants

Participants were from India, recruited through Amazon Mechanical Turk™ (AMT), a trusted and commonly used data collection service (e.g., Snow et al., 2008). These participants aimed to improve English summary writing through this experiment. They were randomly assigned into one of three conditions (formal, informal, and mixed) and completed a 3-hour experiment for a $30 monetary compensation. Ninety-three participants completed the experiments at home or any place where they could access a computer and the Internet. There was an uneven number of participants in each condition due to unidentified technical issues. Table 1 displays participants' demographic information in each condition.

Table 1. Demographic Information of Participants.

		Condition			
		Formal	Informal	Mixed	Total
Gender (N)	Male	21	20	21	62 (66.7%)
	Female	8	9	14	31 (33.3%)
	Total	29	29	35	93
Age	Mean (SD)	33.17 (9.71)	30.83 (6.09)	33.46 (9.62)	32.55 (8.69)
Year of English Learning	Mean (SD)	15.90 (8.44)	17.31 (6.88)	16.43 (10.16)	16.54 (8.64)

Table 2. Information of eight texts: One example of text order.

Sessions	Text Structures	Text ID	Topics	Words	Formality	FKGL
Testing	Compare/contrast	1	Butterfly and Moth	255	0.12	8.6
		2	Hurricane	222	0.20	9.4
	Cause/effect	3	Floods	230	0.47	9.2
		4	Job Market	240	0.62	10.9
Training	Compare/contrast	5	Walking and Running	399	0.18	8.9
		6	Kobe and Jordan	299	0.14	9.2
	Cause/effect	7	Effects of Exercising	195	0.28	9.1
		8	Diabetes	241	0.64	11.7

Note. FKGL = Flesch-Kincaid Grade Level.

Reading materials

Eight short expository texts (see Table 2), ranging from 195 to 399 words, were selected and modified from CSAL AutoTutor lessons (Li & Baer, 2019). These texts, measured by the Coh-Metrix formality, tended to be formal, ranging from 0.12 to 0.64, equivalent to the formality of textbooks of science and social studies above grade 8 to the early years of college (Graesser, McNamara, et al., 2014). Their Flesch-Kincaid Grade Level (FKGL) showed the equivalent levels of text difficulty, ranging from 8.6 to 11.7 grade levels.

Four texts were compare/contrast texts that connected ideas by comparing or contrasting two things or persons and displaying their similarities and differences (Meyer, 2003). Four were cause/effect texts that presented a causal relationship (Meyer, 2003). These two text structures are frequently used in prior studies and considered as the medium level to teach and learn text structures, with compare/contrast texts being much easier than cause/effect texts (Hebert et al., 2016).

Two compare/contrast texts and two cause/effect texts were randomly selected from eight texts for the testing session (i.e., pretest and posttest), with the balanced 4×4 Latin-square design to control for order effects (Edwards, 1951). The remaining two compare/contrast texts and two cause/effect texts were used for the training session and the same Latin-square design was applied to control for order effects (see Table 2). Figure 1 displays the number of reading materials in three sessions: pretest, training, and posttest.

Training session

The training session included a mini-lecture and practices for each of the four texts. Two computer agents interactively presented a mini-lecture on two instructional components of text structures, focusing on identifying signal words and using text maps (Hebert et al., 2016). Signal words indicate and help identify the structure of texts. The use of signal words generates more logical and well-organized texts, which increases coherence and is, therefore, likely to facilitate comprehension (Wijekumar et al., 2013). Table 3 displays some signal words that indicate the

IES Grant: R305C120001

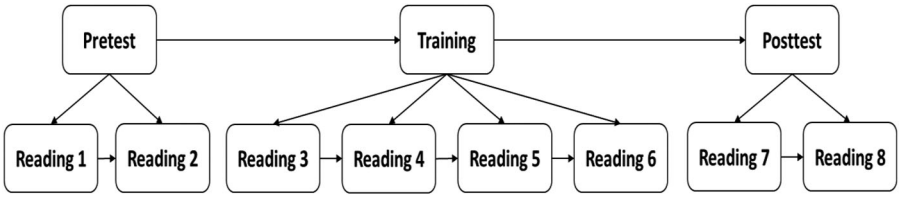

Figure 1. Reading materials on pretest, training, and posttest.

Table 3. Examples of signal words in compare/contrast and cause/effect text structures.

Structures	Categories	Signal Words
Compare/Contrast	Similarities	like, alike, similar, resembles, just as, both, have in common, share, resemble, the same as
	Differences	unlike, differ, on the one hand, in contrast, on the contrary, however, but, in contrast, whereas, in comparison, in opposition
Cause/Effect	Causes	because, since, for the purpose of, if/then, the reason, due to, because of, begin with, when/then,
	Effects	as a result, result in, cause, lead to, consequence, thus, this is why, in order to, so, in explanation, therefore, consequently, effect of

relationships of similarities and differences for the compare/contrast texts and causes and effects for the cause/effect texts.

Agents also utilized text maps (see Figure 2) to demonstrate the presence of the logic relationships of these two text structures. These structures include the usage of dual channels, the visual presentation through graphic organizers and the auditory instruction. These multisensory channels (e.g., non-verbal and verbal channels) benefit learners' comprehension during reading (Paivio, 2017).

After the mini-lectures, two agents interacted with participants and guided them to practice these skills in each text. Each text involved three types of practices: 1) multiple-choice items, 2) summary writing, and 3) summary evaluating.

Through five multiple-choice items, participants practiced identifying the text structure (1 item), finding the main ideas (1 item), and distinguishing the important information from the supporting information (3 items) with the guidance of two agents. After each item, agents gave feedback based on the quality of responses and provided hints if participants gave incorrect answers. Participants had a second chance to modify their responses after they received the hints. Table 4 displays an example of interactions among two agents and a human learner during the training session.

The second practice involved summary writing. Participants were required to write a short summary with 50–100 words. Agents did not provide feedback on the accuracy of written summaries or on the language use in written summaries due to the lack of the accuracy of automated summary scoring.

In the third practice, agents required participants to evaluate the quality of summaries written by themselves and another three summaries, each written by a "peer" agent. These peer summaries were in fact written by one graduate student, an English native speaker, who followed the rubrics for scoring summaries (see Table 5). Another 20 undergraduate students, who were also English native speakers, evaluated the quality of these peer summaries. Results showed that these summaries were assigned into the correct level of quality: good, medium, and poor. The order of three peer summaries was randomly assigned to each text.

Agents did not provide feedback on self-ratings because self-rated summaries were not automatically scored, so the accuracy of self-rating could not be immediately assessed. Agents,

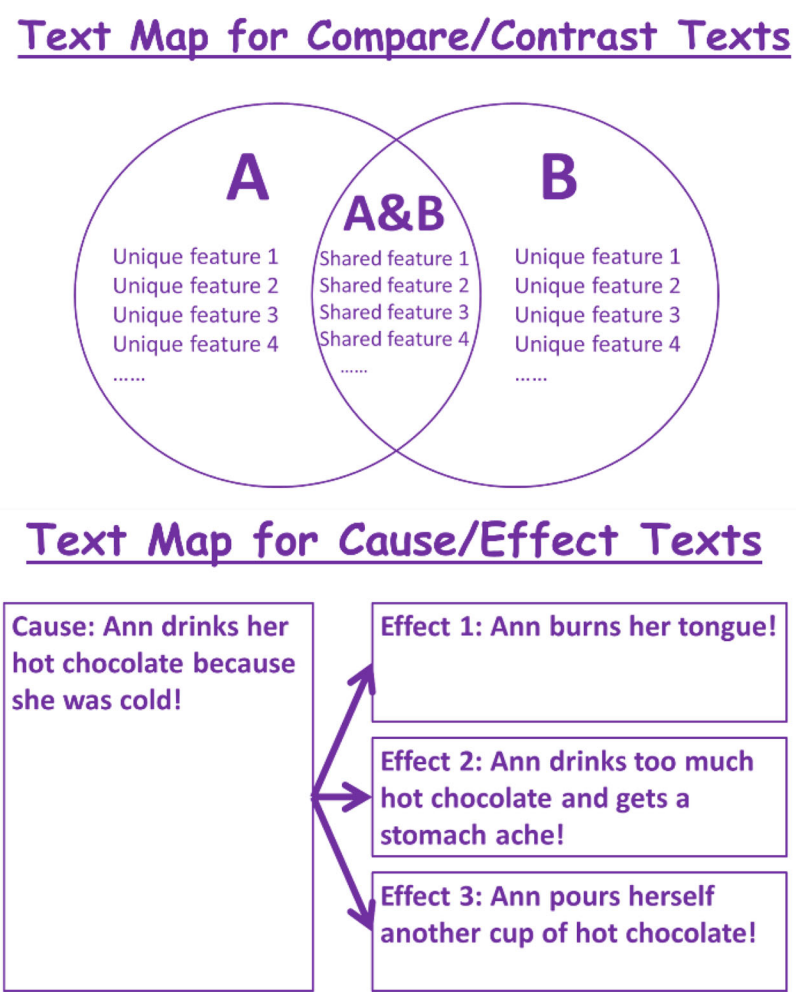

Figure 2. Examples of text maps for compare/contrast and cause/effect texts.

Table 4. An example of trialog that follows expectation & misconception-tailored dialogue (EMT).

Cristina (Teacher Agent): Grace (Participant), which answer summarizes the causes of diabetes? [Main Question]
Grace *(Human learner)*: (*Click*) Diabetes is caused by the excessive fatigue, weight loss, and excess thirst, eye problems, heart problem, and body sores. [*First Trial: Wrong Answer*]
Cristina: Jordan (*Student Agent*), what do you think of this answer? [*Ask Jordan*]
Jordan: This is the correct answer. [*Jordan's Incorrect Response*]
Cristina: This answer shows many symptoms of diabetes. They aren't the causes of diabetes. They occur with diabetes. [*Elaboration*]
Cristina: We should find what causes diabetes from the text. [*Hint*]
Cristina: Try again. I will repeat the question. Grace, could you tell us which answer summarizes the causes of diabetes? [*Repeat Question*]
Grace: (*Click*) Diabetes is caused by the pancreas that makes insufficient insulin or cells that respond to insulin abnormally. [*Second Trial: Correct Answer*]
Cristina: Grace, you got it right! Jordan, you did not give the right answer! [*Feedback*]
Cristina: The third answer shows how people get diabetes. We can find this information from the text. [*Wrap-up*]
Jordan: I see. The text points out two reasons. One is pancreas. Another is cells. The third answer sums up this information. So the third answer is correct. [*Wrap-up*]

IES Grant: R305C120001

Table 5. Rubrics for scoring summary.

Categories	High (2 points)	Medium (1 points)	Low (0 point)
Topic Sentence	The summary begins with a clear topic sentence that states the main idea.	The summary has a topic sentence that touches upon the main idea.	The summary does not state the main idea.
Content Inclusion & Exclusion	Major details are stated economically and arranged in a logical order. No minor or unimportant details or reflections are added.	Some but not all major details are stated and not necessarily in a logical order. Some minor or unimportant details or reflections are added.	Few major details are stated and not necessarily in a logical order. Many minor or unimportant details or reflections are added.
Mechanics and Grammar	There are few or no errors in mechanics, usage, grammar or spelling.	There are some errors in mechanics, usage, grammar or spelling that to some extent interfere with meaning.	There are serious errors in mechanics, usage, grammar or spelling, which make the summary difficult to understand.
Signal Words	Uses the clear and accurate signal words to connect information.	Uses several clear and accurate signal words to connect information.	Uses several clear signal words to connect information.

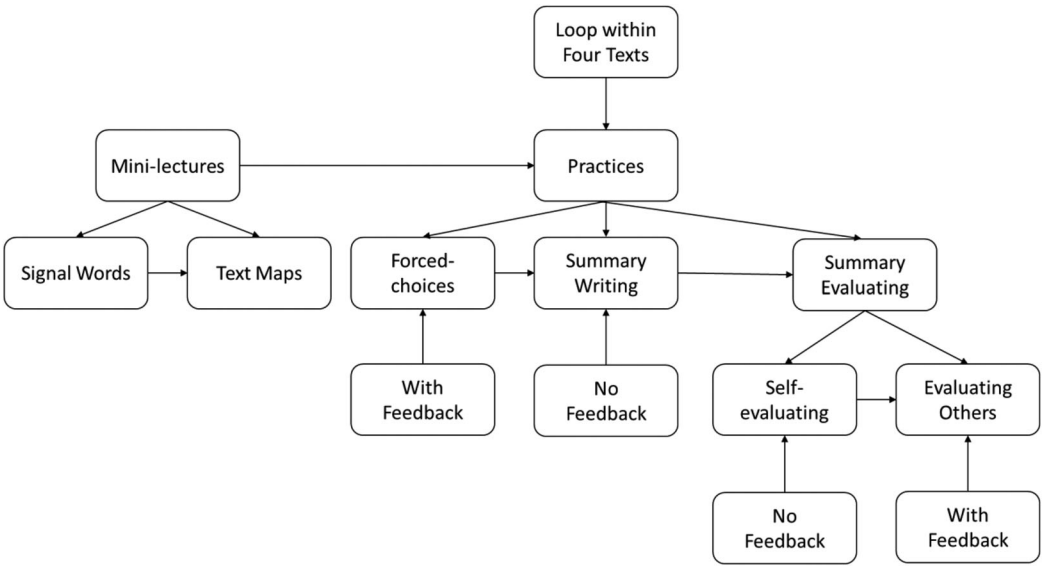

Figure 3. Instructional components during the training session.

however, provided feedback on peer-ratings, including analysis of the quality of peer summaries according to the scoring rubrics and the approach to improving the quality of rated summaries. Figure 3 briefly summarizes the instructional structure in the training session. Figure 4 displays a screenshot of the interface when two computer agents guided the participants to answer the multiple-choice questions.

Test sessions

The pretest and posttest procedures were the same as the training session except for removing the instruction of text structures and five multiple-choice questions. Moreover, agents did not give feedback to the following questions: summary writing, self-rating, or peer-rating.

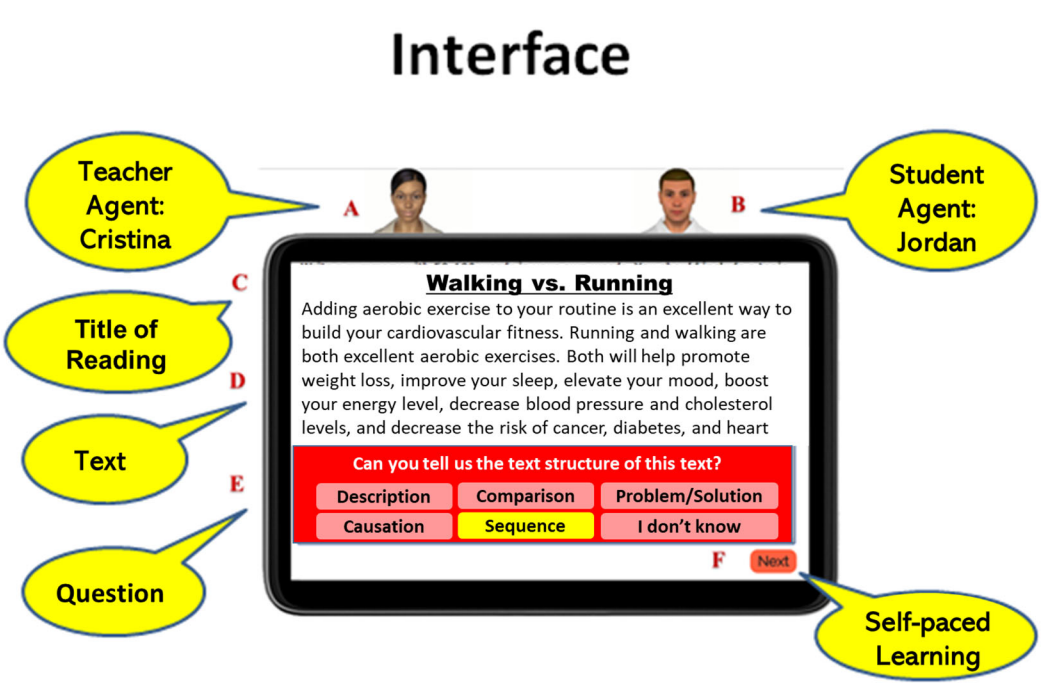

Figure 4. A screenshot for the interface during the training session.
Note. It consisted of (A) the teacher agent, Cristina (female), (B) the student agent, Jordan (male), (C) the title of the reading text, (D) the text presented with the scroll down button, (E) the multiple-choice question for participants to choose an answer during training or input text-box for participants to enter and submit their summaries, and (F) the self-paced next button.

Manipulation

During the design of discourse processing, one expert generated conversations of two agents in formal and informal languages, respectively, following the discourse mechanism, called *expectation & misconception-tailored dialogue* (EMT dialogue) (Graesser, Keshtkar, et al., 2014; Graesser, Li, et al., 2014). EMT dialogue starts when the tutor asks a student a challenging question, then anticipates a particular correct answer. This process typically involves multiple conversational moves, following a five-step tutoring frame:

1. The tutor asks a question;
2. The student gives an initial answer;
3. The tutor gives a short feedback (e.g., positive, negative, or neutral);
4. The tutor interacts with the student and guides the student to reach the expectation (correct answer) through pump–hint–prompt–assertion dialogue moves, including *pumps* (e.g., "anything else"), *hints* (e.g., "have you noticed how authors compared differences of these two basketball players?"), or *prompts* (e.g., "what benefits the patients with diabetes?"); and
5. The tutor provides an assertion as a contribution when the student fails to give a correct answer.

Table 4 displays an example of conversations that reflects the EMT mechanism and tutoring frame. The dialogue move categories are annotated in brackets-with-italics. *Cristina* was the teacher agent, who always had the ground truth; and *Jordan* was the student agent, whose performance was always a bit lower than the human learner. *Grace* was an active human learner, not merely a vicarious observer because she needed to determine her answer based on two agents' suggestions. This dialogue design facilitated learning and engagement (Li et al., 2015).

IES Grant: R305C120001

Table 6. Guidance that determines agent language formality and some examples.

Text Level	Formal Discourse	Informal Discourse
Word	Less familiar words (e.g., visualize) Words with more syllables (e.g., enable)	More familiar words (e.g., show) Words with less syllables (e.g., can)
Syntax	Complex sentences (e.g., the reason why … is that) Complex modifiers (e.g., relationships between causes and effects)	Simple sentences (e.g., perhaps …) Simple modifiers (e.g., causal relationships)
Referential Cohesion	Content words overlapping (e.g., as illustrated in the flow chart (flow chart is repeated; see Table 7)	Less content words overlapping (e.g., for example instead of repeating flow chart, see Table 7)
Deep Cohesion	More connectives (e.g., furthermore, in addition; see Table 7) More causal words (e.g., lead to, consequence, cause, effect)	Less or no connectives (e.g., see the same sentences in Table 7). Less causal words (e.g., cause, effect)
Genre	Use third-person pronouns (e.g., he) Nominalization (e.g. development)	Use first- and second-person pronouns (e.g., you) Verbs (e.g., develop)

Besides EMT dialogue, the expert also followed the multilevel theoretical framework (Graesser & McNamar, 2011) and the pragmatics-based framework (Snow & Uccelli, 2009) to generate conversations of two agents in formal and informal languages, respectively. The guidance that determines language formality consists of five primary text levels: word (e.g., familiar vs. unfamiliar words, words with less or more syllables), syntax (e.g., simple vs. complex sentences and modifiers), referential cohesion (e.g., less or more content words overlapping), deep cohesion (e.g., less or more connectives and causal words), and genre (e.g., first- and second- vs. third-person pronouns, nominalization vs. verbs). Table 6 displays the guidance and some examples for better understanding the mechanism of the development of agent language in formal and informal discourse.

Another expert examined conversations and confirmed that the language of these conversations was natural and appropriate. The formal and informal conversations were then assigned to the teacher agent (Cristina) and student agent (Jordan). Thus, three conditions were generated: formal (for both Cristina and Jordan), informal (for both Cristina and Jordan), and mixed (formal for Cristina and informal for Jordan). The conversations in each condition were evaluated by the Coh-Metrix formality scores (Graesser, McNamara, et al., 2014). The means of agents' formality were −0.37, 0.12, and 1.02 for the informal, mixed, and formal conditions, respectively (Li & Graesser, 2017, 2020). The agents' formality in each condition represents a different level of formality, ranging from informal, to medium, to formal (Graesser, McNamara, et al., 2014). Table 7 displays an example of formal conversations and informal conversations when agents introduced the functions of signal words for cause/effect texts. We did not provide an example of mixed conversations, because the mixed language was generated by combining Cristina's formal language and Jordan's informal language.

Measures

Quality of summaries

Participants were required to write a 50–100-word summary for each of eight informational texts by stating a topic sentence to specify the main idea, providing important information, and employing signal words to explicitly express their ideas. It should be noted that they practiced these three skills during intervention. While the intervention focused on the instruction of signal words through mini-lectures, agents also guided participants to understand these signal words as a function of summarization strategies. Specifically, signal words interrelate main ideas logically, distinguish important information from unimportant information, and consequently enhance

Table 7. Examples of formal and informal discourse.

Formal discourse:
Cristina: Furthermore, a flow chart enables to explicitly visualize the relationships between causes and effects. This graph illustrates the cause that Ann drinks hot chocolate and a series of effects.
Jordan: Suppose that the reason why Ann drinks hot chocolate is that she is cold and needs to warm herself up.
Cristina: Thus, drinking hot chocolate leads to a series of consequences. As illustrated in the flow chart, her tongue is burned, her stomach aches, and she drinks another cup. To sum up, this section introduces some signal words to determine the relationships between similarities and differences in the comparison texts, as well as the relationships between causes and effects, delivered in the causation text. In addition, the text maps enable people to explicitly visualize these relationships.
Informal discourse:
Cristina: A flow chart can best show the relationships between causes and effects. The cause in this example is that Ann drinks hot chocolate.
Jordan: Perhaps Anne is cold, so she drinks hot chocolate to get warm.
Cristina: Yes. Drinking hot chocolate causes many effects. For example, her tongue is burned. Her stomach aches. She drinks another cup. To sum up, some signal words can tell how things are similar or different in the comparison texts. Some words can tell the causal relationships. The flow chart clearly shows these relationships.

efficient processing of text comprehension and improve strategic knowledge about how to use them in different text structures during reading practice (e.g., Meyer and Poon, 2001; Meyer et al., 2018). Thus, the rubric for summary grading included the aforementioned three dimensions. The rubric also includes grammar, given that prior studies on text structure strategy in the ITS included semantic grammar in the rubric (e.g., Meyer et al., 2011). To better compare our findings with prior studies, we added grammar to our rubric. Another reason why we included grammar in the rubric is that we assume that it is possible that training on words and comprehension strategies can have a side effect of improving the syntax in writing. Consequently, summaries were scored by human raters based on a holistic summary writing quality rubric with four dimensions: (1) topic sentence, (2) content inclusion and exclusion, (3) signal words of text structures, and (4) grammar and mechanics (see Table 5; Li & Graesser, 2017, 2020). Each dimension was assessed on a scale of 0–2 points, with 0 for the absence of target knowledge, 1 for the partial presence of knowledge, and 2 for the complete presence of knowledge. Total summary scores ranged from 0 to 8.

Four English-native experts (1 male and 3 females) participated in the three rounds of training for summary scoring after discussing each dimension in the rubrics and then graded three summaries of good, medium, and poor quality, respectively for each text. Each rater graded 32 summaries that were randomly selected from eight texts and then discussed disagreements until an agreement was reached. The average interrater reliabilities for the three training sets reached the threshold (Cronbach $a = .82$). After training, each rater graded summaries for two source texts.

Summary ratings

After participants submitted their summaries, they were asked to rate the quality of their summaries on a 6-point scale: 1 = very bad; 2 = bad; 3 = undecided, but guess bad; 4 = undecided, but guess good; 5 = good; 6 = very good. The quality of rating was measured by the differences between the self-rating scores and the expert-rating scores. Experts graded summary quality using a 0–8 range, so we converted 0–8 range into 1–6 to align with the scale of self-rating scores. Then, we subtracted expert-rating scores from self-rating scores and obtained the absolute differences between self-rating and expert-rating scores without the consideration of negative values for under-rating or positive for over-rating. The smaller scores, therefore, indicated the better performance of self-rating. After they rated their own summaries, participants were asked to rate another three summaries assumed to be written by peers. The same measure of self-rating accuracy was applied to peer-rating accuracy.

IES Grant: R305C120001

Table 8. Means (standard deviations) of Ipre- and post-intervention scores.

Measuring Instrument	Text Structure	Formal		Informal		Mixed		Total	
		Pretest	Postest	Pretest	Postest	Pretest	Postest	Pretest	Postest
Quality	Cause/Effect	3.86(1.35)	3.84(1.45)	4.14(1.38)	4.45(1.38)	3.69(1.59)	3.79(1.84)	3.88(1.45)	4.02(1.60)
	Compare/Contrast	3.99(2.01)	4.28(2.05)	4.21(1.70)	4.86(1.85)	3.54(2.08)	3.97(1.74)	3.89(1.95)	4.35(1.89)
	Total	3.93(1.45)	4.06(1.45)	4.17(1.28)	4.66(1.23)	3.61(1.52)	3.94(1.54)	**3.89(1.43)**	**4.20(1.44)**
Self-Rating	Cause/Effect	1.64(1.00)	1.56(1.00)	1.67(0.95)	1.42(1.01)	1.79(1.11)	1.70(1.39)	1.71(1.02)	1.57(1.16)
	Compare/Contrast	1.76(1.38)	1.55(1.03)	1.67(1.22)	1.82(1.15)	2.23(1.22)	1.61(1.41)	1.91(1.29)	1.66(1.22)
	Total	1.70(1.04)	1.56(0.75)	1.67(0.86)	1.62(0.81)	2.01(0.99)	1.64(1.30)	**1.81(0.97)**	**1.61(1.00)**
Peer-Rating	Cause/Effect	1.14(0.61)	0.61(0.45)	0.94(0.54)	0.88(0.49)	1.01(0.53)	0.67(0.51)	1.03(0.56)	0.72(0.50)
	Compare/Contrast	1.10(0.55)	0.64(0.48)	0.84(0.54)	0.59(0.41)	1.00(0.66)	0.63(0.46)	0.98(0.59)	0.62(0.45)
	Total	1.12(0.46)	0.63(0.38)	0.89(0.44)	0.74(0.33)	1.01(0.51)	0.66(0.44)	**1.01(0.48)**	**0.67(0.39)**
Writing Time (Minutes)	Cause/Effect	8.54(2.92)	6.27(2.95)	8.09(3.39)	6.72(3.40)	7.81(3.57)	6.50(3.51)	8.12(3.30)	6.50(3.28)
	Compare/Contrast	8.71(2.97)	7.00(3.36)	8.61(3.87)	7.87(2.54)	8.21(3.53)	5.61(2.38)	8.49(3.45)	6.75(2.89)
	Total	8.63(2.52)	6.64(2.83)	8.35(2.86)	7.30(2.57)	8.01(3.13)	6.03(2.53)	**8.31(2.85)**	**6.62(2.66)**

Note. Scores on pretest and posttest are bolded in the Total row and Total column. Scores within each agent formality condition on pretest and posttest are in the Total row and the columns of Formal, Informal, and Mixed.

Formality of summaries

Summaries written by participants were analyzed by the text analysis tool Coh-Metrix (3.0) to obtain the five primary Coh-Metrix components: narrativity, word concreteness, syntactic simplicity, referential cohesion, and deep cohesion (Graesser, McNamara, et al., 2014). We reversed the first three component scores (non-narrativity, word abstractness, and syntactic complexity) to be consistent with the latter two and computed the formality scores with the means of five components, where the higher scores represented more formal summaries.

Results

We performed the mixed-effects models to answer two research questions: (1) Does conversational agents' formality (informal, mixed, and formal) during instruction affect learning of text structures? and (2) Does the conversational agents' formality during instruction affect participants' language use when they write summaries? Mixed-effects models combine fixed and random effects and allow for the assessment of the influence of the fixed effects on dependent variables after accounting for any extraneous random effects. The fixed effects were agent formality (informal, mixed, and formal; a between-subjects factor), a repeated measure of time (pretest and posttest), and their interaction. Participants ($N = 93$) were used as the random effect. This study aimed to examine whether agent language affects learning outcomes and language use after the training, so analyses focused on the differences in learning performance and language use between the pretest and the posttest.

The dependent variables in relation to learning performance included the quality of summaries, accuracy of self-rating, accuracy of peer-rating, and summary writing time. The dependent variables in relation to language use in written summaries included formality of summaries and its underlying five Coh-Metrix components: word abstractness, syntactic complexity, referential cohesion, deep cohesion, and non-Narrativity. The scores of all dependent variables were aggregated mean scores of two text structures (cause/effect and compare/contrast) on pretest and posttest, respectively. Therefore, text structure was not able to be included in the analyses for aggregated data. To further examine whether the findings were consistent in each text structure, we split the original dataset into two subsets based on text structure: one was a cause/effect pre-post subset, and another was a compare/contrast pre-post subset. All significance testing was conducted with an alpha level of .05 with Bonferroni correction for multiple analyses. Cohen's d was computed as an appropriate effect size (Cohen, 1988). The following section reports the results for the aggregated dataset and then briefly talks about results for split datasets.

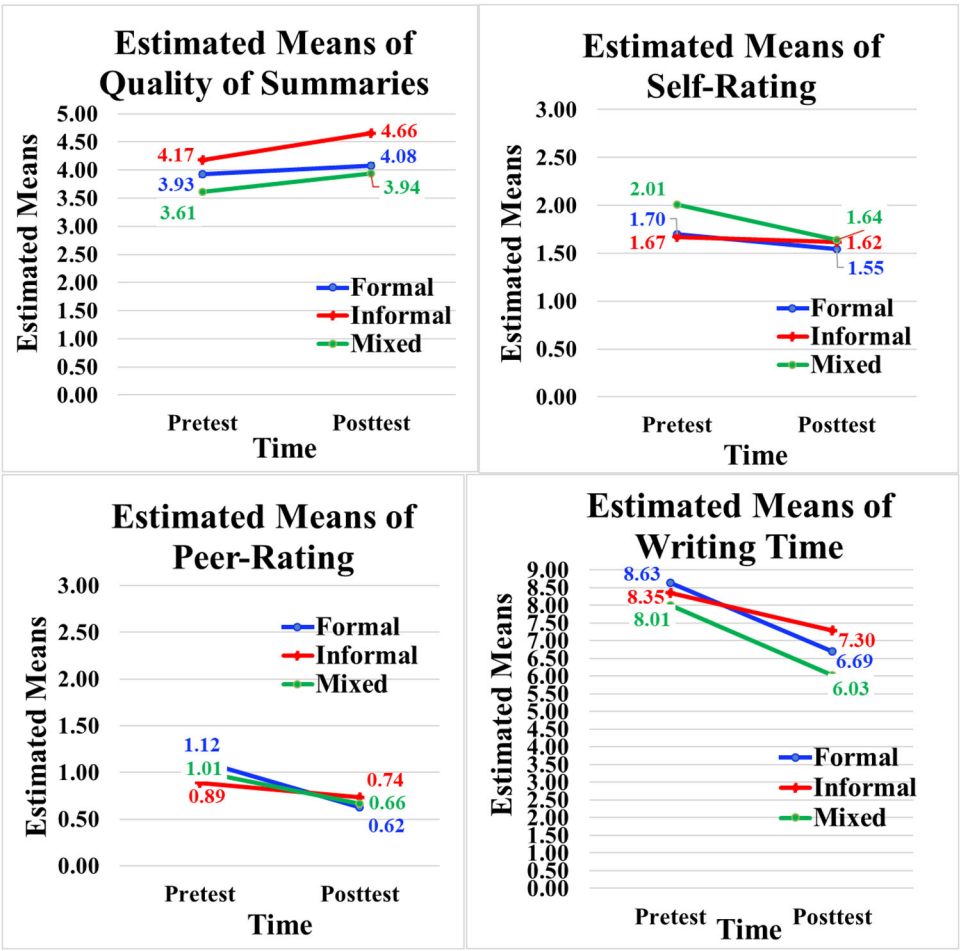

Figure 5. The estimated means of pre- and post-intervention scores.

Learning performance in aggregated dataset

Table 8 displays the mean scores and standard deviations of learning performance on pretest and posttest in each agent formality condition. Figure 5 presents the estimated means of learning performance in each condition.

Quality of written summaries

The mixed-effects linear regression model with the quality of written summaries as the dependent variable showed a significant main effect of time, $F(1, 89.66) = 4.45$, $p = .038$, but did not yield any significant differences in conditions, $F(2, 90.13) = 2.17$, $p = .121$ or in the two-way interaction of time × condition, $F(2, 89.65) = 0.36$, $p = .698$. Further pairwise analyses for time indicated that the participants wrote significantly higher quality summaries on posttest ($M = 4.20$, $SD = 1.44$) than pretest ($M = 3.89$, $SD = 1.43$), Cohen's $d = .18$.

Accuracy of self-rating

The same analysis on differences between self-rating and expert-rating did not show a significant main effect of time ($F(1, 92.56) = 3.16$, $p = .079$), condition ($F(2, 92.98) = 0.59$, $p = .557$), or two-way interaction of time × condition ($F(2, 92.54) = 0.80$, $p = .455$).

IES Grant: R305C120001

Table 9. Means (standard deviations) of language use in three agent formality conditions within text structures on pretest and posttest.

Language Use	Text Structure	Formal		Informal		Mixed		Total	
		Pretest	Postest	Pretest	Postest	Pretest	Postest	Pretest	Postest
Formality	Cause/Effect	0.57(0.58)	0.41(0.51)	0.55(0.52)	0.51(0.44)	0.49(0.49)	0.32(0.44)	0.53(0.52)	0.41(0.46)
	Compare/Contrast	0.16(0.47)	0.21(0.58)	0.30(0.56)	0.15(0.48)	0.17(0.62)	0.26(0.61)	0.21(0.55)	0.21(0.56)
	Total	0.36(0.43)	0.31(0.40)	0.43(0.40)	0.33(0.35)	0.33(0.41)	0.28(0.39)	0.37(0.41)	0.31(0.38)
Word	Cause/Effect	−0.80(0.91)	−0.93(0.87)	−0.39(1.22)	−0.53(1.09)	−0.52(1.31)	−1.15(1.56)	−0.57(1.17)	−0.89(1.25)
Abstractness	Compare/Contrast	−1.30(1.54)	−0.85(1.44)	−1.43(1.47)	−0.90(1.55)	−1.05(1.81)	−1.05(1.39)	−1.25(1.62)	−0.94(1.45)
	Total	−1.05(0.86)	−0.89(0.84)	−0.91(0.99)	−0.71(1.01)	−0.79(0.94)	−1.08(1.20)	−0.91(0.93)	−0.91(1.04)
Syntactic	Cause/Effect	0.58(1.24)	0.08(1.50)	0.39(0.99)	−0.08(1.40)	0.14(1.33)	−0.30(1.78)	0.35(1.21)	−0.11(1.57)
Complexity	Compare/Contrast	0.57(1.39)	0.40(1.08)	0.91(1.29)	0.40(1.12)	0.36(1.00)	0.36(1.01)	0.60(1.23)	0.38(1.06)
	Total	0.57(1.17)	0.24(1.12)	0.65(0.98)	0.16(0.91)	0.25(0.96)	0.05(1.06)	0.48(1.04)	0.14(1.03)
Referential	Cause/Effect	0.24(1.06)	0.04(1.16)	0.24(1.24)	−0.13(1.07)	−0.19(1.29)	0.13(1.09)	0.08(1.21)	0.02(1.10)
Cohesion	Compare/Contrast	0.52(1.34)	0.68(1.30)	1.38(1.57)	0.64(1.45)	0.70(1.20)	1.17(1.49)	0.86(1.40)	0.85(1.43)
	Total	0.38(0.93)	0.36(0.97)	0.81(1.08)	0.26(0.94)	0.26(1.04)	0.64(1. 01)	0.47(1.04)	0.43(0.98)
Deep	Cause/Effect	2.28(2.24)	2.05(2.24)	2.01(1.58)	2.39(2.21)	2.35(1.63)	2.01(1.96)	2.22(1.81)	2.14(2.11)
Cohesion	Compare/Contrast	−0.17(1.50)	−0.24(1.98)	−0.46(1.51)	−0.58(1.36)	−0.34(1.56)	−0.15(1.94)	−0.32(1.51)	−0.31(1.78)
	Total	1.06(1.41)	0.91(1.58)	0.78(1.01)	0.91(1.22)	1.00(1.14)	0.88(1.53)	0.95(1.19)	0.90(1.44)
Non-Narrativity	Cause/Effect	0.53(0.76)	0.81(0.76)	0.50(0.52)	0.92(0.65)	0.68(0.86)	0.90(0.87)	0.58(0.73)	0.88(0.76)
	Compare/Contrast	1.18(0.79)	1.04(0.76)	1.11(0.76)	1.17(0.64)	1.21(1.22)	0.97(0.76)	1.17(0.96)	1.05(0.72)
	Total	0.86(0.59)	0.93(0.65)	0.80(0.49)	1.04(0.45)	0.94(0.79)	0.93(0.65)	0.87(0.64)	0.96(0.59)

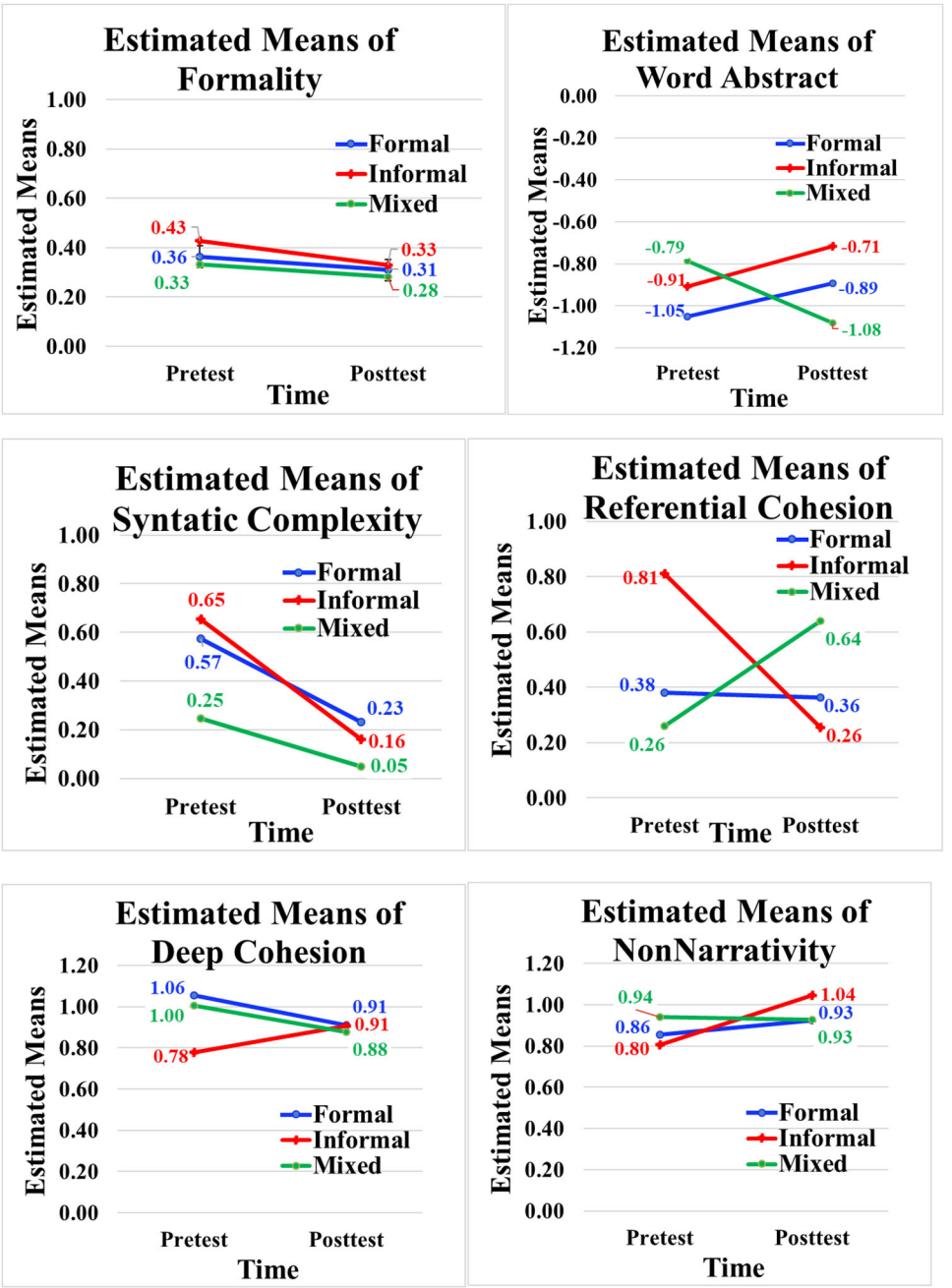

Figure 6. The estimated means of language use.

Accuracy of peer-rating

The same analysis on differences between peer-rating and expert-rating showed a significant main effect of time, $F(1, 93.04) = 31.17$, $p < .001$. Participants rated peer summaries more accurately on posttest than on pretest because their peer-rating scores were significantly lower on posttest ($M = 0.67$, $SD = 0.39$) than on pretest ($M = 1.01$, $SD = 0.48$), Cohen's $d = .61$ (see Table 8). Results did not show a significant main effect of condition ($F(2, 93.12) = 0.26$, $p = .773$), or two-way interaction ($F(2, 93.02) = 2.58$, $p = .081$).

IES Grant: R305C120001

Writing time

The analysis for writing time also showed a significant main effect of time on writing, $F(1, 92.45) = 39.00$, $p < 0.001$. Students spent less time writing summaries on posttest ($M = 6.62$ min, $SD = 2.66$) than pretest ($M = 8.31$ min, $SD = 2.85$), Cohen's $d = .49$ (see Table 8). Results did not show a significant main effect of condition ($F(2, 93.00) = 1.01$, $p = .367$) or two-way interaction ($F(2, 92.43) = 1.27$, $p = .285$).

Use of language in written summaries in aggregated dataset

Table 9 displays the mean scores and standard deviations of language use on pretest and posttest in each agent formality condition. Figure 6 presents the estimated means of formality and five Coh-Metrix components in each condition.

Formality of written summaries

The mixed-effects model for formality did not show a significant main effect of time ($F(1, 92.53) = 1.39$, $p = .242$), condition ($F(2, 92.53) = 0.53$, $p = .593$), or two-way interaction ($F(2, 92.51) = 0.06$, $p = .938$.

Five Coh-Metrix components of written summaries

The mixed-effects model did not show a significant main effect of time, condition, or their two-way interaction between time and condition for word abstractness, deep cohesion, or non-narrativity, with all p values larger than .05.

In contrast, results showed a significant main effect of time on syntactic complexity, $F(1, 92.93) = 7.91$, $p = .006$, but did not show a significant main effect of condition ($F(2, 93.26) = 1.04$, $p = .359$) or their two-way interaction ($F(2, 92.92) = 0.50$, $p = .606$). Further pairwise comparison analyses showed that participants used less complex sentence structures when they wrote summaries on posttest ($M = 0.14$, $SD = 1.03$) than pretest ($M = 0.48$, $SD = 1.04$), Cohen's $d = .27$ (see Table 9). Results for referential cohesion showed no significant main effect of time ($F(1, 92.92) = 0.19$, $p = .665$) or condition ($F(2, 92.87) = 0.40$, $p = .675$), but a significant two-way interaction, $F(2, 92.90) = 3.48$, $p = .035$. Further pairwise comparison analyses showed that participants used significantly less referential cohesion in their summaries on posttest ($M = 0.26$, $SD = 0.94$) than pretest ($M = 0.81$, $SD = 1.08$) in the informal condition ($F(1, 92.34) = 4.47$, $p = 0.037$, Cohen's $d = .43$). This pattern did not occur in the formal ($F(1, 92.98) = 0.01$, $p = .946$) or mixed condition ($F(1, 92.34) = 2.54$, $p = .115$).

Analyses for split subsets: cause/effect vs. compare/contrast

To further examine whether the findings in the aggregated dataset were consistent with those in each text structure, we performed the same series of mixed-effects models on the two text structure subsets, cause/effect and compare/contrast respectively. Tables 8 and 9 display the mean scores and standard deviations of learning performance and language use on pretest and posttest in each agent formality condition and within each text structure, respectively. Due to the limited space, we only report results that were significant with p values less than .05.

Compare/contrast data subset

Results in the compare/contrast subset were consistent with those in the aggregated dataset in relation to learning performance. Specifically, participants wrote better quality summaries ($M_{Pretest} = 3.89$, $SD_{Pretest} = 1.95$; $M_{Posttest} = 4.35$, $SD_{Posttest} = 1.89$, $F(1, 93.02) = 4.79$, $p = 0.031$, Cohen's $d = .19$), more accurately rated peer summaries ($M_{Pretest} = 0.98$, $SD_{Pretest} = 0.59$; $M_{Posttest} = 0.62$,

Table 10. Summary of the significant analysis results.

	Aggregate Dataset			Compare/Contrast Dataset			Cause/Effect Dataset		
Measure	Time	Condition	Interaction	Time	Condition	Interaction	Time	Condition	Interaction
Quality of summary	X			X					
Accuracy of self-rating									
Accuracy of peer-rating	X			X			X		X
Writing time	X			X			X		
Formality of summary									
Word abstractness									
Syntactic complexity	X						X		
Referential cohesion			X						
Deep cohesion									
Non-narrativity							X		

Note. X means that the result was significant, with p less than .05. Time refers to the difference between pretest and posttest. Condition refers to the difference among conditions: formal, mixed, and informal. Interaction refers to the difference between the time × condition interaction.

$SD_{Posttest} = 0.45$, $F(1, 93.18) = 20.58$, $p < 0.001$, Cohen's $d = .54$), and spent less time writing summaries ($M_{Pretest} = 8.49$, $SD_{Pretest} = 3.45$; $M_{Posttest} = 6.75$, $SD_{Posttest} = 2.89$, $F(1, 92.12) = 21.65$, $p < 0.001$, Cohen's $d = .43$) on posttest than on pretest. Results did not show any significant main effects of time or condition, or interactions between time and condition for language use, including formality and its five primary components: word abstractness, syntactic complexity, referential cohesion, deep cohesion, or non-narrativity, which was inconsistent with results in aggregated dataset.

Cause/effect data subset

Slightly different findings were shown in the cause/effect texts for both learning performance and language use. The consistent findings included: participants more accurately rated peer summaries ($M_{Pretest} = 1.03$, $SD_{Pretest} = 0.56$; $M_{Posttest} = 0.72$, $SD_{Posttest} = 0.50$, $F(1, 90.66) = 18.21$, $p < 0.001$, Cohen's $d = .47$) and spent less time writing summaries ($M_{Pretest} = 8.12$, $SD_{Pretest} = 3.30$; $M_{Posttest} = 6.50$, $SD_{Posttest} = 3.28$, $F(1, 92.77) = 18.93$, $p < 0.001$, Cohen's $d = .40$) on posttest than on pretest. No significant main effect of time was found in the quality of summaries as the aggregated data and compare/contrast subset revealed. However, a significant interaction between time and condition was found in peer-rating, $F(2, 90.67) = 3.42$, $p = .037$. Further pairwise analyses within each agent formality condition showed that the differences between peer-rating and expert-rating scores were significantly smaller on posttest than pretest when they interacted with the agents who spoke the formal ($M_{Pretest} = 1.14$, $SD_{Pretest} = 0.61$; $M_{Posttest} = 0.61$, $SD_{Posttest} = 0.45$, $F(1, 91.30) = 16.72$, $p < .001$, Cohen's $d = .77$) and mixed discourse ($M_{Pretest} = 1.01$, $SD_{Pretest} = 0.53$; $M_{Posttest} = 0.67$, $SD_{Posttest} = 0.51$, $F(1, 91.02) = 8.36$, $p = .005$, Cohen's $d = .53$). This pattern was not found in the informal condition, $F(1, 89.71) = 0.19$, $p = .661$.

Moreover, a significant main effect of time was found in syntactic complexity, which was consistent with the aggregated dataset. Participants used less complex syntactic structures in summaries for cause/effect texts ($M_{Pretest} = 0.36$, $SD_{Pretest} = 1.21$; $M_{Posttest} = -0.11$, $SD_{Posttest} = 1.57$, $F(1, 91.93) = 7.41$, $p = 0.008$, Cohen's $d = .29$), which was consistent with the findings shown in the aggregated dataset. Unlike the aggregated and compare/contrast subset, participants wrote more non-narrative summaries on posttest than on pretest, ($M_{Pretest} = 0.58$, $SD_{Pretest} = 0.73$; $M_{Posttest} = 0.88$, $SD_{Posttest} = 0.76$, $F(1, 93.51) = 6.52$, $p = 0.012$, Cohen's $d = .33$).

Discussion and conclusions

In this study, we investigated the impact of agent formality on summary writing in an ITS from the perspectives of participants' learning and language use. Ninety-three adults from India

IES Grant: R305C120001

participated in the three-hour experiment through AMT and learned the structures of compare/contrast and cause/effect texts in one of three conditions: (1) where both the teacher agent and the student agent spoke the formal discourse, (2) where both agents spoke the informal discourse, or (3) where the teacher agent spoke the formal discourse and the student agent spoke the informal discourse, called the mixed discourse. We answered two research questions: (1) Does the conversational agents' formality during instruction affect learning of text structures? and (2) Does the conversational agents' formality during instruction affect the use of language?

We used four measures (i.e., quality of summary writing, accuracy of self-rating, accuracy of peer-rating, and writing time) in three data sets (i.e., aggregated, compare/contrast, cause/effect) to answer the first question, with 12 measures in total. We utilized six measures (formality, word abstractness, syntactic complexity, referential cohesion, deep cohesion, and non-narrativity) in the same three data sets to answer the second research questions, with 18 measures in total.

Table 10 summarizes the significance of the results in each data set for an explicit visualization.

Impact of agent formality on learning

Seven of 12 measures for learning showed a significant difference between pretest and posttest, but no difference was found among three conditions or among the interaction between time and condition. These results imply that agent language formality did not affect learning of summary writing when measured by quality of summary, accuracy of peer-rating, or writing time in both aggregated and compare/contrast data set as well as accuracy of peer-rating in cause/effect data set (see Table 10). Four of these measures showed no significant difference in time, condition, or their interaction, including quality of summary writing in cause/effect data set and accuracy of self-rating in three data sets. These results suggest that agent language formality did not promote learners to improve self-rating performance or summary writing performance. Only one of 12 measures showed a significant interaction, which is accuracy of peer-rating. This finding indicates that language formality enhanced students' performance on peer-rating from pretest to posttest within the constriction to cause/effect texts when agents spoke the formal and mixed discourse.

However, this result has to be taken cautiously for the following two reasons. First, the validity of this measure is not strong because participants rated the quality of summaries with different criteria as experts did. Experts used four dimensions to grade summaries: topic sentence, content inclusion/exclusion, signal words of text structures, and grammar and mechanics, with each dimension assigned a score 0–2 points, with the total of 0–8 points. Participants rated peers' summaries from very bad to very good, with the total of 1–6 points. During the instruction and feedback, the agents used the four dimensions that experts utilized to provide comments and feedback on participants' written summaries, but it is unclear whether participants used these dimensions to rate their and peers' summaries. Second, the significant difference in accuracy of peer rating was only found for cause/effect texts, but not for compare/contrast texts or both texts. Therefore, it may not be generalizable until we have more empirical evidence in further studies that use the same criteria for expert rating and self-/peer- rating and include more text structures such as problem/solution, description, and sequence.

In sum, the answer to the first research question is that agent language formality (formal, informal, mixed) does not affect learning of summary writing (11 out of 12 measures) with regards to text structure according to the evidence revealed in our study. These findings are inconsistent with prior findings that agents who spoke the informal discourse better enhanced learning than agents who spoke the formal discourse (Ginns et al., 2013; Lin et al., 2020; Moreno & Mayer, 2000, 2004; Reichelt et al., 2014; Riehemann & Jucks, 2018). Our findings did not reveal that participants benefited from the agents who spoke the informal discourse. Instead, they benefited from the agents who spoke the formal and mixed discourse in peer-rating in cause/effect

text structures. One reason why participants did not improve the quality of summaries or self-rating when they interacted with the formal or mix agents is likely that participants only received feedback on the performance of peer-rating, but not on self-rating or quality of written summaries due to the lack of real-time assessment of summaries. These findings further suggest that immediate feedback provided by pedagogical computer agents in learning cause-effect text structures could facilitate learning more than no feedback (Li et al., 2018). Moreover, when participants and agents have no shared background on how to generate a good summary or evaluate a summary, the agents should use more formal language to elaborate on the new information more explicitly until there is common understanding (Clark, 1996).

The findings on quality of summaries and summary writing time in the present study were inconsistent with our prior studies (2017, 2020). The present study compared the performance on pretest and posttest, and investigated the effectiveness of the intervention that agents provided in three styles of language. Previous studies focused on the impact of agent language on the performance of each summary or of summaries at each stage (i.e., two summaries on pretest, four during training, and another two on posttest) and reported the effective results in relation to three styles of language. The present study reported performance in terms of learning gains on pretest vs. posttest, learning gains within each condition on pretest vs. posttest, and learning gains within each condition in each text structure on pretest vs. posttest in split subsets.

The present study advances prior studies in a number of ways. First, the present study provided empirical evidence that the complex measures of agent language at multiple levels of language and discourse (Li & Graesser, 2017, 2020) showed inconsistent results with studies in which agent language was manipulated at the personal-pronoun level (Ginns et al., 2013; Lin et al., 2020; Moreno & Mayer, 2000, 2004; Reichelt et al., 2014; Riehemann & Jucks, 2018). Agents' formal, mixed, or informal speech did not promote learning in summary writing. More empirical evidence is needed to examine why the agent language measured by complex measures yields different findings than when it is measured by personal pronouns.

Second, the present study extended the investigation of the learning of science (e.g., Moreno & Mayer, 2000, 2004), psychology (Reichelt et al., 2014; Riehemann & Jucks, 2018), or summary writing (Li & Graesser, 2017, 2020) to a higher level of learning, where participants applied acquired knowledge of text structure and summary writing to assess the quality of their own summaries and peer summaries. As participants become assessors, they are required to engage in a more thoughtful understanding of the rating processes and criteria to enable a critical analysis of the work of others, and for identification of misunderstandings or gaps in thinking (Searby & Ewers, 1997). We did not find an impact of agent language on self-rating, possibly because self-rating is more challenging than peer-rating as people tend to want to reflect a good image of themselves (Shrauger & Osberg, 1981), which may lead them to over-estimate their own work compared to the work of others (see Table 8).

Third, our findings for summary writing were inconsistent with the study by Gámez and Lesaux (2015), who found that students' reading comprehension performance had a significant correlation with teachers' use of sophisticated, academic vocabulary. One possible reason is that Gámez and Lesaux's study used multiple-choice question items to assess comprehension, while our study used summary writing, self-rating, and peer-rating, as higher levels of cognitive assessment. As mentioned before, the improvement found in peer-rating implies that for higher-level cognitive learning, real-time feedback and scaffolding are critical along with the teacher/agent language. Another reason is due to different measures of teacher language: Gámez and Lesaux's study focused on sophisticated, academic vocabulary, whereas our study focused on more comprehensive levels of language. Yet another possible reason is that the effect of agent language in computer-assisted environments is likely to differ from traditional classroom settings, which needs further evidence to confirm.

IES Grant: R305C120001

To sum up, these findings indicate that agent discourse formality does not affect learning of summary writing when agent language was measured by multiple textual levels of language/discourse and when performance was measured by higher level cognitive tasks, including summary writing, self-rating, and peer-rating.

Impact of agent formality on language use in written summaries

Fourteen of 18 measures showed no significant difference in time, condition, or their interaction (see Table 10). These findings indicated that student language did not vary from pretest to posttest and agent language formality did not affect student use of language. These nonsignificant measures include formality, word abstractness and deep cohesion in three data sets, syntactic complexity in the compare/contrast data set, referential cohesion in both compare/contrast and cause/effect data sets, and non-narrativity in aggregated and compare/contrast data sets. Three of 18 measures showed a significant difference between pretest and posttest, but no difference was found among conditions or the Time × Condition interaction. These results imply that participants' language use varied from pretest to posttest in syntactic complexity in both aggregated and cause/effect data sets and in non-narrativity in cause/effect data set, but agent language formality did not affect participants' use of language in these levels. Only one of 18 measures, i.e., referential cohesion, showed a significant interaction in the aggregated data set. This result suggested that agent informal language facilitated learners to use lower referential cohesion on posttest than prestest.

In sum, the answer to the second research question is that agent language formality (formal, informal, mixed) does not affect learners' use of language in summary writing (14 out of 18 measures), but a small variation in language use was found in syntactic complexity and non-narrativity (3 out of 18) from pretest to posttest. However, agent language formality did not affect these results. Agent language formality only affected one measure: referential cohesion. Specifically, participants tended to have lower referential cohesion when they interacted with the agents who spoke more informal discourse when comparing posttest to pretest. This pattern was not found in formal or mixed conditions. Therefore, the informal agent language resulted in summaries with lower referential cohesion.

This finding partially confirms the claim that agent language affects student language. Prior studies revealed that students' use of vocabulary was positively correlated with teachers' use of sophisticated vocabulary and complex syntax (Gámez & Lesaux, 2012). Our study provided more empirical evidence through a causal experiment that participants' use of less referential cohesion was affected by the agents' informal language, such as less repetition of the content words (see Table 7), because agents' informal discourse relies more on world knowledge to fill in what the textbase does not explicitly articulate. The reason that we did not find the impact of agent language on participant language at the levels of word, syntactic complexity, deep cohesion, or genre, is likely due to the repetition of content words being easier to acquire implicitly as compared to other language features when explicit language instruction is lacking. The reason why agent formal and mixed language did not improve participants' use of more referential cohesion is likely that participants tend to choose a much simpler and familiar writing style between the agents' formal language and their own informal language when they are not required to write in a formal language. Further investigation is needed to confirm whether the explicit instruction of the formal language during the intervention and the clear requirement of formal writing in the instruction will enhance students' use of the formal language for academic writing when they interacted with the agents who spoke formally.

Moreover, we found that participants used less complex syntactic complexity in both the aggregated dataset and the cause/effect subset when comparing posttest to pretest. This pattern was not found in compare/contrast texts. These findings indicate that text structures affected

participants' use of syntactic complexity in written summaries. The further examination of the syntactic complexity of the source texts revealed that the syntactic complexity in cause/effect source texts was only 0.23 on average, less than that in compare/contrast source texts, 0.61. The lower syntactic complexity in source texts led to participants use of less complex syntactic structures. Perhaps, the participants imitated the source texts and used simple sentences more than the complex sentences in writing. This explanation could also be used to explain the phenomenon that in cause/effect texts participants used a more non-narrative style on posttest than on pretest. A further examination of the use of non-narrativity in source texts revealed that cause/effect source texts had higher non-narrativity, −0.35 on average, than that in compare/contrast source texts, −1.96. More empirical evidence is needed to better understand how the language use in source texts impacts students' language use in their writing and which affects participant language more, agent language or source text language.

One limitation of the study was that we did not investigate the effect of text difficulty, text interest, or other text characteristics, such as domain-specific versus domain-general texts. These factors may affect learning and language use along with agents' formality. Another concern was that the intervention was short and lasted about three hours. The long-term intervention with a much larger dosage, such as 30 hours of intervention with AutoTutor that varies language style, is needed in future studies to further manifest whether interventions could potentially influence the outcome measures. In the future, the intervention may be allotted into different time periods to see whether the same pattern occurs. Moreover, a future study may include one agent who uses a mixed discourse whose formality falls between formal and informal discourse, as opposed to having the two discourses used by two distinct agents.

Another limitation is that we did not ask participants to use the same criteria for summary ratings as expert raters used even though they were guided in the training and provided feedback on these dimensions: how to identify topic sentences, main ideas, and important information through the multiple-choice questions when they read texts during the intervention. While we assumed that participants knew about these criteria, it is unclear what criteria participants used to rate these summaries. Further studies are needed to give learners explicit instruction for the rating criteria that experts use. The same rating criteria provided to both experts and learners will increase the validity of the summary rating accuracy measure and help investigate whether agent formality facilitates more accurate rating of summaries.

The third limitation of this study is that participants were adult English learners in Indian countries. Empirical evidence is needed to determine whether the findings are universal by including participants who are native English speakers with different English levels, such as upper primary school students, secondary school students, or college students.

References

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Denton, P. (2013). *The power of our words: Teacher language that helps children learn*. Center for Responsive Schools, Inc.

Edwards, A. L. (1951). Balanced Latin-square designs in psychological research. *The American Journal of Psychology*, *64*(4), 598–603. https://doi.org/10.2307/1418200

Fillmore, L. W., & Snow, C. E. (2003). What teachers need to know about language. In C. A. Adger, C. E. Snow, & D. Christian (Eds.), *What teachers need to know about language* (pp. 7–54). Center for Applied Linguistics.

Galloway, E. P., & Uccelli, P. (2015). Modeling the relationship between lexico-grammatical and discourse organization skills in middle grade writers: Insights into later productive language skills that support academic writing. *Reading and Writing*, *28*(6), 797–828. https://doi.org/10.1007/s11145-015-9550-7

Gámez, P. B., & Lesaux, N. K. (2012). The relation between exposure to sophisticated and complex language and early-adolescent English-only and language minority learners' vocabulary. *Child Development*, *83*(4), 1316–1331. https://doi.org/10.1111/j.1467-8624.2012.01776.x

IES Grant: R305C120001

Gámez, P. B., & Lesaux, N. K. (2015). Early-adolescents' reading comprehension and the stability of the middle school classroom-language environment. *Developmental Psychology*, *51*(4), 447–458. https://doi.org/10.1037/a0038868

Ginns, P., Martin, A. J., & Marsh, H. W. (2013). Designing instructional text in a conversational style: A meta-analysis. *Educational Psychology Review*, *25*(4), 445–472. https://doi.org/10.1007/s10648-013-9228-0

Graesser, A. C., Keshtkar, F., & Li, H. (2014). The role of natural language and discourse processing in advanced tutoring systems. In T. Holtgraves (Ed.), *The Oxford Handbooks of Language and Social Psychology* (pp. 491–509). Oxford University Press.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, *23*(5), 374–380. https://doi.org/10.1177/0963721414540680

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topicsin Cognitive Science*, *3*, 371

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, *115*(2), 210–229. https://doi.org/10.1086/678293

Hebert, M., Bohaty, J. J., Nelson, J. R., & Brown, J. (2016). The effects of text structure instruction on expository reading comprehension: A meta-analysis. *Journal of Educational Psychology*, *108*(5), 609–629. https://doi.org/10.1037/edu0000082

Kalinowski, E., Gronostaj, A., & Vock, M. (2019). Effective professional development for teachers to foster students' academic language proficiency across the curriculum: A systematic review. *AERA Open*, *5*(1), 1–23. https://doi.org/10.1177/2332858419828691

Li, H. & Baer W. (2019) Scaffolding adult learners's reading strategies in the intelligent tutoring system. In K. Millis, D. Long, J.P. Magliano, & K. Wiemer (Eds.) *Deep Comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp. 166–179). Abingdon, UK: Taylor and Francis.

Li, H., Gobert, J., Graesser, A. C., & Dickler, R. (2018). Advanced educational technology for science inquiryassessment. *Policy Insights from the Behavioural and Brain Sciences*, *5*(2), 171.

Li, H., & Graesser, A. C.(2020) Impact of conversational formality on the quality and formality of written summaries. In I. Bittencourt, M. Cukurova, K. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education. AIED* 2020. Lecture Notes in Computer Science (Vol. 12163, pp. 321–332). Springer.

Li, H., & Graesser, A. C. (2017) Impact of pedagogical agents' conversational formality on learning and engagement. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.) *Artificial Intelligence in Education. AIED* 2017. Lecture Notes Computer Science (Vol. 10331, pp. 188–200). Springer.

Li, H., Cheng, Q., Yu, Q., & Graesser, A. C. (2015). The role of peer agent's learning competency in trialogue-based reading intelligent systems. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education: Lecture Notes in Computer Science* (Vol. 9112, pp. 694–697). Springer International.

Lin, L., Ginns, P., Wang, T., & Zhang, P. (2020). Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education*, *143*, 103658.

Lucero, A. (2014). Teachers' use of linguistic scaffolding to support the academic language development of first-grade emergent bilingual students. *Journal of Early Childhood Literacy*, *14*(4), 534–561. https://doi.org/10.1177/1468798413512848

Meyer, B. J. F. (2003). Text coherence and readability. *Topics in Language Disorders*, *23*(3), 204–224.

Meyer, B. J., & Poon, L. W. (2001). Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology*, *93*(1), 141–159.

Meyer, B. J. F., Wijekumar, K., & Lei, P. (2018). Comparative signaling generated for expository texts by 4th–8th graders: variations by text structure strategy instruction, comprehension skill, and signal word. *Reading and Writing*, *31*(9), 1937–1968. https://doi.org/10.1007/s11145-018-9871-4

Meyer, B. J., Wijekumar, K. K., & Lin, Y. C. (2011). Individualizing a web-based structure strategy intervention for fifth graders' comprehension of nonfiction. *Journal of Educational Psychology*, *103*(1), 140–168.

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology*, *92*(4), 724–733. https://doi.org/10.1037/0022-0663.92.4.724

Moreno, R., & Mayer, R. E. (2004). Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology*, *96*(1), 165–173. https://doi.org/10.1037/0022-0663.96.1.165

Paivio, A. (2017). A dual coding perspective on imagery and the brain. In *Neuropsychology of visual perception* (pp. 203–216). Routledge.

Reichelt, M., Kämmerer, F., Niegemann, H. M., & Zander, S. (2014). Talk to me personally: Personalization of language style in computer-based learning. *Computers in Human Behavior*, *35*, 199–210. https://doi.org/10.1016/j.chb.2014.03.005

Riehemann, J., & Jucks, R. (2018). "Address me personally!": On the role of language styles in a MOOC. *Journal of Computer Assisted Learning*, *34*(6), 713–719. https://doi.org/10.1111/jcal.12278

Searby, M., & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the School of Music, Kingston University. *Assessment & Evaluation in Higher Education*, *22*(4), 371–383.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin*, *90*(2), 322–351. https://doi.org/10.1037/0033-2909.90.2.322

Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson, & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112–133). Cambridge University Press.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). *Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks*. In M. Lapata. Association for Computational Linguistics.

Wijekumar, K. K., Meyer, B. J., & Lei, P. (2013). High-fidelity implementation of web-based intelligent tutoring system improves fourth and fifth graders content area reading comprehension. *Computers & Education*, *68*, 366–379.